

# Decoding Al series

#### Session 2

Unleashing the power of Large Language Models

Part 1 – Direct interactions

#### Our awesome speaker



Gil Fernandes Al Solutions Engineer One**point** 

Welcome

#### From Roots to Fruits

• Journey to create value from LLMs





Agenda

Insights from previous webinar	<u>C</u>
Interacting with LLMs	<u>C</u>
Natural Language Processing	1
Extracting structured data	
Live demo	<u> </u>
Dark data	<u> </u>
Credits	
Thank you for joining	

<u>04</u> 07 <u>16</u> <u>33</u> <u>46</u> <u>49</u>

<u>55</u>



### Insights from previous webinar



parameter paral provide para

Insights from previous webinar

# LLMs are a technological milestone with shortcomings

LLM achievements

Great understanding of language

Trained on variety of NLP tasks



Models can be trained for different scenarios

Models are adapted to behave like humans

Insights from previous webinar

# LLMs are a technological milestone with shortcomings

LLM shortcomings

Limited reasoning

Hallucinations

Knowledge cut-offs



Missing sources

Refusals



# How we interact with LLMs (prompting)

Ad hoc

One-shot

Positive and negative examples

Context-based





## Audience poll



# How we interact with LLMs (constraints)

Without memory

With memory

- Custom instructions
- Recent chat history



#### Customize ChatGPT

#### Custom Instructions 🕕

What would you like ChatGPT to know about you to provide better responses?

I am a software developer based in London who likes to code in Python, but also likes languages like Rust. I am very interested in LLMs and machine learning.

157/1500

How would you like ChatGPT to respond?





# How we interact with LLMs (possibilities)

**Direct interaction** 

Workflow based

Agents

#### Agent ensembles

Workflow based, agents and agent ensembles are covered in an upcoming TechTalk





Source: Onepoint



### How we interact with LLMs (possibilities)

**Direct interaction** 

Workflow based

Agents

#### Agent ensembles

Workflow based, agents and agent ensembles are covered in an upcoming TechTalk







### How we interact with LLMs (possibilities)





Source: Onepoint



# How we interact with LLMs (possibilities)

**Direct interaction** 

Workflow based

Agents

#### Agent ensembles

Workflow based, agents and agent ensembles are covered in an upcoming TechTalk





Source: Onepoint

### Audience insights





preserved a parel provide par-

# LLMs as powerful language processors

- LLMs trained to be chatbots or follow instructions.
- Both tasks require impressive NLP skills.



17

### NLP skills



Translation

Summarisation

Single or multi-turn dialogue

Zero-shot classification

# LLMs as powerful language processors

Translation

- One language to another
- Most LLMs are multilingual, but not equally skilled in all languages





Source: Gemini standard



# LLMs as powerful language processors

Summarisation

- One language to another
- Great at extracting the essence of text



Can you please summarize the following text in two sentences:

Foundation models (FMs), or large models pretrained on massive data then adapted for downstream tasks, have emerged as an effective paradigm in modern machine learning. The backbone of these FMs are often sequence models, operating on arbitrary sequences of inputs from a wide variety of domains such as language, images, speech, audio, time series, and genomics (Brown et al. 2020; Dosovitskiy et al. 2020; Ismail Fawaz et al. 2019; Oord et al. 2016; Poli et al. 2023; Sutskever, Vinyals, and Quoc V Le 2014). While this concept is agnostic to a particular choice of model architecture, modern FMs are predominantly based on a single type of sequence model: the Transformer (Vaswani et al. 2017) and its core attention layer (Bahdanau, Cho, and Bengio 2015) The efficacy of self-attention is attributed to its ability to route information densely within a context window, allowing it to model complex data. However, this property brings fundamental drawbacks: an inability to model anything outside of a finite window, and quadratic scaling with respect to the window length.

An enormous body of research has appeared on more efficient variants of attention to overcome these drawbacks (Tay, Dehghani, Bahri, et al. 2022), but often at the expense of the very properties that makes it effective.

As of yet, none of these variants have been shown to be empirically effective at scale across domains. Recently, structured state space sequence models (SSMs) (Gu, Goel, and Ré 2022; Gu, Johnson, Goel, et al. 2021) have emerged as a promising class of architectures for sequence modeling. These model scan be interpreted as a combination of recurrent neural networks(RNNs) and convolutional neural networks (CNNs), with inspiration from classical state space models (Kalman 1960). This class of model scan be computed very efficiently as either a recurrence or convolution, with linear or near-linear scaling in sequence length. Additionally, they have principled mechanisms for modeling long-range dependencies (Gu, Dao, et al. 2020) in certain data modalities, and have dominated benchmarks such as the Long Range Arena (Tay, Dehghani, Abnar, et al. 2021). Many flavors of SSMs (Gu, Goel, and Ré 2022; Gu, Gupta, etal. 2022; Gupta, Gu, and Berant 2022; Y. Li et al. 2023; Ma et al. 2023; Orvieto et al. 2023; Smith, Warrington, and Linderman 2023) have been successful in domains involving continuous signal data such as audio and vision (Goel et al. 2022; Nguyen, Goel, et al. 2022; Saon, Gupta, and Cui 2023). However, they have been less effective at modeling discrete and information-dense data such as text.

Foundation models (FMs), primarily based on Transformer sequence models, are effective in handling diverse data types but face limitations like finite context windows and quadratic scaling. Recently, structured state space sequence models (SSMs) have emerged as a promising alternative, offering efficient computation and strong performance on long-range dependencies, although they struggle with discrete, information-dense data like text.

Source: ChatGPT 4o

20

# LLMs as powerful language processors

Single or multi-turn dialogue

- Ask questions and receive replies
- The model can also have "memory" regarding previous questions





Source: Onepoint

21

# LLMs as powerful language processors

Zero-shot classification

 Define specific categories in one prompt and then ask the model to categorise a piece of text according to the presented categories



Can you please categorise the following text according to these categories "arts", "sports", "IT", "science" and output the probabilities from 0 to 1 for each of the categories, Make sure that the probabilities you define add up to 1. AutoGPT, ChatDEV, AutoGEN and metaGPT are some of the top multiagent libraries released this year alone. Microsoft has recently launched an exciting piece of software which is called TaskWeaver and there is so much talk about it in AI spaces. specifically designed to allow users to turn their ideas into code with just a few commands. TaskWeaver is not just another tool; It's a complex system that interprets your commands, converts them into code, and performs tasks with precision. In this video, we will discuss what is TaskWeaver, what are the features of Taskweaver, the differences between Autogen and TaskWeaver and how TaskWeaver can be used to create a super AI Agent. The given text has been categorized with the following probabilities for each category: \$ • Arts: 0.0 • Sports: 0.0 • IT: 1.0 • Science: 0.0

This indicates that the text is entirely related to the IT category. [>-]

Source: GPT 4o

22

### More NLP skills



Sentiment Analysis

Named entity recognition

Part of Speech (POS) tagging

23

#### LLMs as powerful language processors

Sentiment analysis

Ask LLMs like Mixtral-8x22B Instruct v0.1 about the sentiment of text and they will tell you



Source: https://api.together.ai/playground/chat/mistralai/Mixtral-8x22B-Instruct-v0.1



Can you please let me know which is the sentiment of this text; whether it is positive, neutral or negative:

Unspooling in dialogue of embarrassing banality, "Rebel Moon" finally buckles beneath the weight of its hackneyed exposition.

24

# LLMs as powerful language processors

Named entity recognition

 Define specific categories in one prompt and then ask the model to categorise a piece of text according to the presented categories



Can you please extract me the names of all locations and of government institutions in this following text:

••

Commuters faced misery on roads and rail networks and people were told to prepare for flooding as the first named storm of 2024 – Henk – swept across parts of southern Britain, bringing winds of up to 80mph (128km/h).

About 2,000 homes in Cornwall experienced power cuts and trees were brought down by the wind. People were asked to stay away from coastal areas hit by the storm and in some areas police advised people not to travel unless absolutely necessary.

The London fire brigade said the storm was likely to hit the capital between 3pm and 6pm and asked people to secure loose objects outside homes, such as decorations, bins or furniture, not to shelter under trees and to watch out for debris when driving.

The Met Office put an amber severe weather warning in place from 10am to 8pm on Tuesday, saying there would probably be some damage to buildings and travel disruption, such as tiles blown from roofs and longer journey times and cancellations, with some bridges and roads likely to close.

\$

Sure, here are the extracted names of locations and government institutions from the given text:

#### Locations:

- 1. Southern Britain
- 2. Cornwall
- 3. Coastal areas
- 4. London

#### Government Institutions:

- 1. London fire brigade
- 2. The Met Office



# LLMs as powerful language processors

Part of Speech (POS) tagging

• Use of LLMs for POS tagging using standard notations



56 ד 56 T	OKENS   107.28 TOKENS/S
	Can you please perform Part of Speech Tagging using the Penn Treebank tagset notation for the following text:
	for the following text.
	"Artificial Intelligence (AI) holds the key to unlocking the immense potential of 'dark'
	company data — a treasure trove often comprising unstructured elements like images,
	sound recordings, and various forms of text. This untapped potential, hidden in plain sight, can revolutionize how we understand and utilize information.
	of Speech Tagging for the given text using the Penn Treebank tagset
otation:	Copy Code
otation: Artificia	Copy Code
Artificia Intellige	Copy Code
otation: Artificia	Copy Code
Artificia Intellige (AI/NNP)	Copy Code
Artificia Intelliger (AI/NNP) holds/VBZ the/DT key/NN	Copy Code
Artificia Intelliger (AI/NNP) holds/VBZ the/DT key/NN to/TO	Copy Code 1/NNP hce/NNP
Artificia Intelliger (AI/NNP) holds/VBZ the/DT key/NN	Copy Code 1/NNP hce/NNP

Source: https://api.together.ai/playground/chat/meta-llama/Llama-3-70b-chat-hf



### More NLP skills



#### Keywords extraction

Language generation

Masked language

27

# LLMs as powerful language processors

Keywords extraction

LLMs are also capable of extracting keywords from text









Source: Gemini standard

28

# LLMs as powerful language processors

Language generation

- LLMs are great at generating text based on instructions
- Emulate different writing styles



Can you please generate a soonet in Shakesperian style on how Yoda deals with the Monday Blues when he comes to the office?

#### Sonnet: Yoda and the Monday Blues

When dawn breaks forth and brings the Monday's plight,The weary soul of Yoda stirs from rest.With furrowed brow and eyes of glowing light,To office bounds, he ventures on his quest.

The week's beginning, fraught with tasks anew, He breathes in deep, the Force within to guide. Though many woes the Monday may construe, In wisdom's cloak, young Yoda shall abide.

To coffee's call he yields his keenest ear, A cup to banish sleepy discontent. Through meditations, clarity draws near, The blues of Monday start their swift descent.

In office halls, his presence shines so bright, For Yoda turns each burden into light.

Source: ChatGPT 4.0



## LLMs as powerful language processors

Masked language

- Filling blanks in text
- A task LLMs excel at, as they were initially trained to perform this task



William Thomson, 1st Baron Kelvin, OM, GCVO, PC, FRS, FRSE (26 June 1824 – 17 December 1907)[7] was a British mathematician, mathematical physicist and engineer \_\_\_\_\_\_ in Belfast. [8] He was the professor of Natural Philosophy at the \_\_\_\_\_\_ of Glasgow for 53 years, when he undertook significant research and mathematical analysis of electricity, was instrumental in the \_\_\_\_\_\_ of the first and second \_\_\_\_\_\_ of thermodynamics,[9] [10] and contributed significantly to the \_\_\_\_\_\_ physics, which was then in its infancy of development as an emerging academic \_\_\_\_\_\_. He received the Royal Society's Copley Medal in 1883 and served as its president from the 1890 to 1895. In 1892, he became the first \_\_\_\_\_\_ to be elevated to the house of \_\_\_\_\_\_. [11]

William Thomson, 1st Baron Kelvin, OM, GCVO, PC, FRS, FRSE (26 June 1824 – 17 December 1907)[7] was a British mathematician, mathematical physicist and engineer \_born\_in Belfast. [8] He was the professor of Natural Philosophy at the \_university\_of Glasgow for 53 years, when he undertook significant research and mathematical analysis of electricity, was instrumental in the \_formulation\_of the first and second \_laws\_of thermodynamics,[9] [10] and contributed significantly to the \_unifying\_physics, which was then in its infancy of development as an emerging academic \_discipline\_. He received the Royal Society's Copley Medal in 1883 and served as its president from the 1890 to 1895. In 1892, he became the first \_scientist\_to be elevated to the house of \_Lords\_. [11]

Source: https://en.wikipedia.org/wiki/Lord\_Kelvin



#### Natural Language Processing NLP business use case

Create prompt to translate, summarise and extract main ideas related to competitiveness and innovation of a text in German



Source: https://claude.ai/ Claude Sonnet 3.5



#### Summary:

The text discusses the importance of Artificial Intelligence (AI) in the BMW Group's strategy and operations. It highlights AI's role in automating complex processes, improving efficiency, and adding value across the entire value chain. The company has implemented a Data & AI Initiative to facilitate knowledge transfer and has over 600 AI use cases. The text also emphasizes BMW's commitment to responsible digitalization, outlined in five guidelines that prioritize human-centric innovation, security, transparency, and privacy.

Main points regarding competitiveness:

- 1. AI implementation across the entire value chain, enhancing efficiency and quality of business processes.
- 2. Over 600 AI use cases, demonstrating extensive integration of AI technologies.
- 3. Holistic and sustainable approach to value creation across the company.
- 4. Focus on responsible use of digital technology, potentially enhancing trust and reputation.

Main points regarding innovation:

- 1. Recognition of AI as a key technology for the 21st century, driving technological advancement.
- 2. Data & AI Initiative promoting rapid networking, knowledge, and technology transfer throughout the company.
- 3. Emphasis on identifying new technologies and conceiving value creation across the entire company.
- 4. Five guidelines for modern information and communication technology, putting people at the center of digital innovations.
- 5. Commitment to high standards for usability and cybersecurity in digital innovations.

Source: https://claude.ai/ Claude Sonnet 3.5

31

## Natural Language Processing NLP use case

#### Create prompt to categorise questionnaire according to self-defined rules.

Determine how confident you are in terms giving advice to a customer based on a sequence of questions and answers that you can find here:

•••

{questions\_answers}

•••

In order to be confident you should know about the difficulties of the customer. You should know about the following:

\* the customer's main problem

\* have some detailed information about his problem. Just having a high-level sentence on the problem of the customer like e.g: "Data Quality" is not enough to be confident.

- \* you should have also knowledge about what is causing the problem.
- \* ideally you know more than one cause behind the main problem of the customer.
- \* you also need some more background information about the technologies used by the customer to be confident.
- you should also know about the data governance strategies of the customer to be able to have a high degree of confidence.

Please use the following classifications to this question about the degree of confidence with which you can give advice:

- "outstanding"

- "high"

- "medium"
- "mediocre"
- "low"

Source: Onepoint



For example, you should report an "outstanding" confidence degree when:

You know the main problem of the customer and the causes well. You also know about data governance aspects related to the customer's organisation.

And you also know the technological landscape of the customer very well.

For example, you should report a "high" confidence degree when:

You know the main problem of the customer and the causes well. You also know about data governance aspects related to the customer's organisation.

The only thing missing is more background information about the technological landscape of the customer.

For example, you should report a "medium" confidence degree when:

You know the main problem of the customer and the causes well.

You miss the background information about the technological landscape of the customer and also about data governance in the customer's company.

For example, you should report a "mediocre" confidence degree when:

You know the main problem of the customer and the cause not well. The information about the causes is very limited You miss the background information about the technological landscape of the customer and also about data governance in the customer's company.

For example, you should report a "low" confidence degree when: You know the main problem of the customer and nothing else. Or you do not even know about the main problem of the user.

32

### Extracting structured data



ALL DESCRIPTION OF A DE

O COM STORE OF A STORE

preserved as parent preserved party

### Types of data



#### Structured data sources

Unstructured data sources

Semi-structured

Extracting Structured Data

Types of data

Structured data

Used in most databases

Variety of algorithms are based on this data



Used to govern the business processes of enterprises



Extracting Structured Data

Types of data

#### Unstructured data sources

Text

- Written by humans
- Other sources serves logs

Images

- Generated by humans
- Other sources –
  CCTV cameras



#### Audio

- Music
- Voice recordings

Video

- Generated by humans
- Automated sources


Types of data

Semi-structured

Sensor data

- Like temperature
- Data generated by planes or cars

Spatial data

Location data 

**GIS** data 



### Rich media

3D models 



## Using LLMs to extract data from text: problem

### LLMs have powerful NLP skills which allow to extract structured data.



38

## Using LLMs to extract data from text: problem

**Problem:** "classical" LLMs are trained either to follow instructions or to act as chatbots. Extracting data required writing unreliable parsers



Extract an ontology. Do not draw it, but create a markdown table of relations with 3 columns in this order: source, target, relation name. And also create a list with 2 columns: the first contains the term name and the second column contains the term definition. the relations should only be mappeded to source, target, relations Use this text as a basis:



#### \$ **Relations** Table

Source	Target	Relation Name
Rory McIlroy	2024 US Open	haunted by
high-profile pundits	Rory McIlroy	comment on
Rory McIlroy's coach	Rory McIlroy	questions recovery
Four birdies	McIlroy's final-round lead	contributes to
Rory McIlroy	fifth major title	in pursuit of
Rory McIlroy	first since 2014	in pursuit of
Rory McIlroy	world No 3	is
Rory McIlroy	bogey three of the last four holes	does

Source: ChatGPT 4o



## Using LLMs to extract data from text: solution

OpenAl released in June 2023 "Function calling"

- Allows the API callers to specify a data schema for a function in your code.
- The output is formatted as JSON.
- The first models with this feature were gpt-3.5-turbo-0613 and gpt-4-0613.





## Using LLMs to extract data from text: solution

Nowadays there are lots of models which support this feature



In terms of open-source models there are more than 220 models which support function calling:

<u>https://huggingface.co/models?other=function+calling&sort=trending</u>





S OpenAl Platform Gemini

# Ontology creation using function call: input

Step 1

Specify message



#### "messages": [

ر {

"role": "system",
"content": "You are an assistant."

#### "role": "user",

"content": "Extract an ontology. Do not draw it, but create a markdown table of relations with 3 columns in this order: source, target, relation name. And also create a list with 2 columns: the first contains the term name and the second column contains the term definition. \nthe relations should only be mappeded to source, target, relations\nUse this text as a basis:\n\n``Four birdies in the space of five holes from the ninth had given McIlroy a two-shot final-round lead in pursuit of his fifth major title and a first since 2014, only for the world No 3 to bogey three of the last four holes. That spell included missed par putts of two feet six inches and three feet nine inches on the 16th and 18th respectively and allowed Bryson DeChambeau to claim his second US Open title with a brilliant par save from a bunker on the last.```\n\n"

Source: ChatGPT 4o



# Ontology creation using function call: input

Step 2

Specify function format



```
"functions": [
       "name": "create_ontology",
       "description": "Creates an ontology from the give text
       containing topic relationships",
       "parameters": {
           "type": "object",
           "properties": {
               "relationships": {
                   "title": "Relationships",
                   "description": "The list of all edges in the
                   questionnaire",
                   "type": "array",
                   "items": {
                        "$ref": "#/definitions/Edge"
            <u>ر (</u>
           "required": [
               "relationships"
           , [
           "definitions": {
               "Edge": {
                   "title": "Edge",
                   "description": "Represents single triplet with
                   source, target and relationship names",
                   "type": "object",
```

Source: ChatGPT 4o



# Ontology creation using function call: output

The incoming JSON message will contain the specified ontology as function arguments which can be fed into a SQL database.



# Ontology creation using function call: output



Source: ChatGPT 40





Database

## Live demo



And the second second second second

personal parts (provide ())

Distance Black / Black

ness balance watch - Bangle Chester

Free AI tool

# Explore your data health with Onepoint D-Well<sup>™</sup>



Embark on your data wellness journey with our free tool, powered by generative AI and the Onepoint Data & Analytics Body of Knowledge™







Scan the QR code or visit <u>www.onepointltd.com/data-wellness</u>



### Free AI tool Onepoint D-Well<sup>™</sup> (Technical concept)









## Dark data



personal parts (provide ())

Distance Black / Black

ness balance watch - Bangle Chester

### Dark data What is dark data?

Dark data is the **vast amount of information** that organisations collect, process, and store during regular business activities, but fail to use for other purposes such as analytics, business relationships, or direct monetisation.





Dark data What is dark data?

It can be used to gain insights to

Optimise business processes

Help with anomaly detection

Contain data which might violate privacy laws





51

## Audience poll



52

#### Dark data

## Strategies to deal with dark data

### Dealing with Dark Data requires a multi-step approach.





## Audience insights







www.perplexity.ai



gemini.google.com



api.together.ai/playground/chat/mi stralai/Mixtral-8x22B-Instructv0.1



api.together.ai/playground/chat/met a-llama/Llama-3-70b-chat-hf





openai.com/index/gpt-4



<u>claude.ai</u>



en.wikipedia.org/wiki/Lord\_Kelvin

## Thank you for joining

Please feel free to contact Gil Fernandes if you have any feedback about the session.



Email at <u>techtalk@onepointltd.com</u>

Connect on LinkedIn <u>www.linkedin.com/in/gil-palma-fernandes</u>

Find Gil's Reflections on AI at <u>medium.com/@gil.fernandes</u>







#### Session 3

# Spotlight on dark data with AI

Tuesday, 27 August 2024 | 11:00 UK | 12:00 CEST

Our awesome speaker



Allan Schweitz Director of Technology -Services





Scan the QR code or visit <u>onepointItd.com/techtalk/spotlight-on-</u> <u>dark-data-with-ai/</u>







#### Session 4

# Unleashing the power of Large Language Models

Part 2 – Workflows and complex interactions

Tuesday, 17 September 2024 | 11:00 UK | 12:00 CEST

Our awesome speaker



Gil Fernandes Al Solutions Engineer One**point** 



**Branches** 

Unleashing the power of Large Language Models -Workflows and complex interactions

Scan the QR code or visit <u>onepointItd.com/techtalk/unleashing-</u> <u>the-power-of-large-language-models-part-2-workflows-and-</u> <u>complex-interactions/</u>



Onepoint is a **boutique**, values-driven, businessoriented technology consultancy.

We architect, prototype, build, and manage data and AI powered solutions. We partner with global clients looking for high-impact, enterprise-grade advice and IT services to realise their most critical digital transformations.

London | Manchester | Pune

### Onepoint

Your trusted companions for the digital journey<sup>m</sup>



onepointltd.com

hello@onepointItd.com





© Copyright 2024. Onepoint Consulting Ltd is a company registered in England. Company registration number 5516457. VAT registration number GB866120039. ISO27001 certified. Certified in the UK as an ethnic minority-owned business.

